DOCUMENT RESUME

ED 269 430 TM 860 260

AUTHOR Green, Kathy E.; Stager, Susan F.
TITLE Effects of Training, Grade Level, and Subject Taught on the Types of Tests and Test Items Used by Teachers.
SPONS AGENCY Wyoming State Dept. of Education, Cheyenne.
PUB DATE Apr 86
NOTE 13p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 17-19, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Course Content; Elementary School Teachers; Elementary Secondary Education; *Instructional Program Divisions; *Measurement Techniques; Multivariate Analysis; Secondary School Teachers; *Teacher Education; Test Construction; *Test Format; *Test Items; Test Use
IDENTIFIERS Wyoming

ABSTRACT

Teachers' use of classroom tests and test items was examined by level of teacher training in tests and measurement, grade taught and subject area taught. A survey form was developed that contained questions about training in tests and measurement, subject areas and grades taught, from what source test items were taken, hours spent in testing-related activities, the percent of students' grades based on test scores, use of six types of test items, and use of five types of tests. A random sample of 555 practicing elementary and secondary Wyoming teachers participated (81 percent response rate)--a sample size adequate for analyses by grade level. Mean frequency of test and item use was calculated by amount of training, by grade level, and by content area. The significance of differences in usage were assessed using multivariate analysis of variance followed by univariate tests. The results of significance tests indicate that there are clear differences in testing techniques used by teachers at different grade levels in different subjects. Differences in test use were found between teachers with two or more tests and measurement courses and teachers with no coursework or one course. (PN)

EFFECTS OF TRAINING, GRADE LEVEL, AND SUBJECT TAUGHT ON THE TYPES OF TESTS AND TEST ITEMS USED BY TEACHERS*

Kathy E. Green, Ph.D.

University of Wyoming

Susan F. Stager, Ph.D.

Indiana University

*Appreciation is expressed to the Wyoming State Department of Education for their support of this research. This paper was presented at the National Council on Measurement in Education Annual Meeting, San Francisco, 1986.

ABSTRACT

Training in tests and measurement, grade level taught, and subject area taught were all found to have significant effects on teachers' use of different types of tests and test items. A random sample of 555 practicing teachers in the State of Wyoming participated (81% response rate). Results suggest that flexibility in testing is enhanced by training in tests and measurement beyond the typical basic course. Results also provide information which may be used in tailoring tests and measurement courses and are discussed in those terms.

Teacher use of tests

Testing in U.S. schools continues to be practiced extensively, though debate continues on its place and its value. Given the widespread use of tests with its potential to help or to hinder, it is essential that the assessments made and the methods or instruments used to make them be both of high quality and appropriate to the situation and purpose. Students' motivation to achieve and their perceptions of the educational system may be damaged by inadequate testing practices at any level. Recently several authors have argued that college training in tests and measurement may not be adequately oriented to what teachers actually need, thus limiting their facility in using appropriate testing techniques (Ebel, 1967; Fennessey, 1982; Gullickson, 1984b; Newman & Stallings, 1982). Gullickson (1984b) calls for the development of strategies to meet teachers' needs but notes that prerequisite to this is simply a description of teachers' testing behavior.

Studies of testing practice in the U.S. have consistently found extensive test use. Carlberg (1981) reported that 15% of class time was devoted to testing. In a survey conducted by Newman and Stallings, (1982), teachers reported spending more than 10% of their time dealing with tests. Gullickson (1982) found that 95% of the teachers he surveyed tested at least biweekly. The estimated average percentage of students' course grades which are based on test scores is 40-50% with a range of 0-100% (Gullickson, 1984b; McKee & Manning-Curtis, 1982; Newman & Stallings, 1982). Tests, thus, are used frequently. But how are they used--and how does their use vary with tests and measurement training, content area, and level taught?

Several studies have been conducted relating flexibility in testing practice to training, grade level, and subject taught. The number of purposes for which tests are used and the number of item types used were found to relate to knowledge of measurement principles (Newman & Stallings, 1982). Those teachers with higher knowledge scores tended to use tests for more purposes and to use more item types. However, Fennessey (1982) found no relationship between training and the number or types of tests used.

Grade level taught has been found to be related to test use and to attitudes toward testing. Fewer tests were found to be given at lower than at higher grade levels (Gullickson, 1982; Yeh et al., 1981) and attitudes toward testing were less positive at the lower grade levels (Tollefson et al., 1985).

Use of item types and evaluation techniques have also been found to vary across grade levels and subject area taught (Chambers, 1982; Gullickson, 1984a). Newman and Stallings (1982), for example, found teachers to use completion items, multiple-choice, matching, true-false, short answer, and essay questions (from most to least frequently). Gullickson (1982) found teachers to use objective item types most, followed by essay items. Use of textbook/teachers' manuals as item sources decreased as grade level increased.

The purpose of this paper was to examine classroom test and item use by level of training, grade taught, and subject area taught. Hypotheses were:

1. There are significant differences in the number of item and test types used among teachers with 0, 1, 2, and 3+ courses in tests and measurement, with flexibility increasing as amount of training increases.

2. There are significant differences in the number of item and test types used among teachers at the elementary, junior high, and senior high levels.
3. There are significant differences in the number of item and test types used among teachers in different subject areas (art/music, English, physical education, social science, mathematics, and science).

METHODS

Instruments

A survey form was developed containing questions about training in tests and measurement, subject areas and grades taught, from what source test items are taken, hours spent in testing-related activities, the percent of students' grades based on test scores, use of six types of test items, and use of five types of tests.

Types of items used was treated both as an aggregate (α=.70) and as six separate variables. Questions asked for frequency of use (1=never, 6=always) of:

- true-false questions
- essay questions
- multiple-choice questions
- short answer questions
- completion questions
- matching questions

Types of tests used was also treated as an aggregate (α=.66) and as five separate variables. Questions assessed the frequency of use of:

- diagnostic tests
- norm-referenced tests
- criterion-referenced tests
- performance tests
- competency tests

Subjects

Our goal was to survey approximately 500 teachers--a sample size adequate to allow analyses by grade level. The size of the sample was based on expectations of a 70% return rate. A systematic random sample was chosen from the State Department of Education list of all licensed educators. During the spring semester, these teachers were sent a letter explaining the nature of the study, a survey form, and a stamped return envelope. A return rate of 55% was obtained from the first mailing. With two follow-ups, a total of 555 replies were received, or 81% of the deliverable envelopes. (Twelve were undeliverable, 4 refused, and 133 did not reply.)

The sample included a greater percentage of females (64%), primarily as a consequence of the over-representation of females among elementary school teachers. The greatest percentage of teachers in the total sample and at each of the three grade levels was in the 30-39 year-old range. The average number of years of teaching experience was 12. All teachers in the sample held bachelors' degrees, with 23% holding masters'. Subject area responsibilities seemed representative of public school teachers: the majority of elementary teachers were responsible for all areas; at the junior and senior high levels the most frequently reported areas were in core subjects (English, math, science, social studies, physical education, art/music). Training in tests and measurement was

consistent across grade levels taught: 27% had no coursework, 47% had one course, 17% had two courses, and 9% had three or more courses.

Analyses

Mean frequency of test and item use were calculated by amount of training, by grade level, and by content area. The significance of differences in usage were assessed using multivariate analysis of variance followed by univariate tests.

RESULTS

Significant multivariate effects of coursework in tests and measurement were found for types of tests used ($F_{15,1287}=5.22$, $p<.01$) but not for use of item types, providing partial support for hypothesis 1. Persons with more coursework reported more frequent use of all types of tests, with major increases occurring between groups with 0-1 and 2-3+ courses.

Hypothesis 2 was supported: There were significant multivariate differences across grade levels taught in both use of different item types ($F_{12,972}=8.85$, $p<.01$) and use of different types of tests ($F_{10,882}=4.10$, $p<.01$). Use of true-false and essay items increased significantly from kindergarten through the 6th grade and continued to increase through high school. Use of multiple-choice, short answer, completion, and matching items increased through grade 4 and then dropped slightly in grades 5 and 6. With the exception of short answer items, differences in use of these item types at upper grade levels were not significant. Use of short answer items increased significantly between the lower and the upper grade levels. Use of diagnostic tests was highest in grades 1-4. Use of other types of tests did not differ significantly across the elementary grade levels. Use of both norm- and criterion-referenced tests decreased

significantly across the upper grade levels while use of performance and competency tests did not. Elementary grade teachers reported a significantly heavier reliance on teachers' manuals as a source of test items than did teachers at upper grade levels and spent significantly less time per week in testing-related activities(4.7 hours per week vs. 7.4 at junior high and 6.9 at senior high).

Table 1 presents the same information broken down by content area. Elementary teachers were excluded from this analysis. Univariate F-statistics and significance levels are noted for each variable in Table 1. Hypothesis 3 was supported: Significant multivariate differences were found for both use of item types ($F_{30,830}=5.36$, $p<.01$) and use of test types ($F_{25,692}=5.00$, $p<.01$).

(Table 1 about here)

DISCUSSION

The purpose of this study was primarily to describe differences across grade levels and subject areas in use of different types of tests and test items. The results of significance tests indicate that there are clear differences in testing techniques used by teachers at different grade levels in different subjects. This result is consistent with those found by Chambers (1982) and Gullickson (1984a). The fact that significant differences exist in testing practices across grade level and content area is not surprising: different testing techniques lend themselves more readily to the assessment of different skills. This study serves to describe and highlight the differences.

At the elementary levels, diagnostic tests are used frequently, tests being developed with the aid of teachers' manuals. Standardized tests are used extensively at this level as well. Techniques for early diagnosis and remediation are essential knowledge for elementary level teachers. Completion

and matching items are used more frequently at this level than other item types. At the middle school level, performance tests and competency tests are used more frequently as are short answer items. However, the entire range of item types, both subjective and objective, comes into play. At the high school level, objective item types as well as essay and short answer are all used. Diagnostic (and standardized) tests are used less frequently. Given the limited amount of time devoted to tests and measurement in college curricula, it is appropriate to emphasize different types of tests and items in course sections offered for prospective elementary, middle school, and high school teachers. Alternatively, tests and measurement instructors may need to demonstrate by concrete example how all test and item types can be useful at all levels. (This information is not provided in detail in the major textbooks.)

Use of test and item types at the high school level varies with area taught. English teachers reported more extensive use of subjective than of objective item types; short answer items were used most frequently by mathematics teachers. In most areas, tests were given to assess performance or achievement more frequently than to diagnose difficulties. Fennessey (1982) argues that tests and measurement training should be focused on the student's curricular area--English, physical education, mathematics--whenever possible as well as being structured to respond to needs of prospective elementary, middle school, and hish school teachers. Such structuring of training would involve both great flexibility on the part of instructors and skill on the part of persons who schedule students.

. The use of tests to determine students' grades varied widely across subject taught, so even though emphasis on test construction seems appropriate for classes of prospective high school teachers, for students in fields such as art and music, instruction in alternative assessment techniques is necessary.

Gullickson (1984a) points out that nontest evaluative techniques (lab reports, papers) are used at all grade levels. In areas such as art, music, dance, physical education, and English where less emphasis is placed on test results than in science and math, it is even more important to instruct prospective teachers in evaluation techniques other than paper-and-pencil tests.

Differences in test use were found between teachers with two or more tests and measurement courses and teachers with no coursework or one course. This suggests that coursework beyond the typical undergraduate introductory course will be needed to effect behavioral change. As noted earlier, testing in U.S. schools is extensive with tests being used frequently by both those with and without formal training. Optimal use of tests requires advanced training. This training may perhaps be provided after the teacher has had a year of more's experience rather than as an undergraduate. The majority of baccalaureate programs in teacher education require completion of one course in tests and measurement. This single course does not seem to have the impact on practice that one might wish for. It is suggested that this course (and accompanying texts) be restructured to provide more information directly pertinent to classroom teachers' needs, perhaps by accompanying a basic volume by a second volume composed largely of concrete examples, or that additional training be provided to allow teachers practice in optimal testing techniques.

REFERENCES

Carlberg, C. (1981). South Dakota study report. Denver, Co.: Midcontinent Regional Educational Laboratory.

Chambers, B. (19' '). Quality control review of teacher-made tests. Paper presented at the American Educational Research Association annual meeting, Montreal. (ERIC ED 227166)

Ebel, R.L. (1967). Improving the competence of teachers in educational measurement. In J. Flynn & H. Garber (Eds.), Assessing behavior: readings in educational and psychological measurement. Reading, MA: Addison-Wesley.

Fennessey, D. (1982). Primary teachers' assessment practices: some implications for teacher training. Paper presented at the annual conference of the South Pacific Association for Teacher Education, Frankston, Victoria, Australia. (ERIC ED 229346)

Gullickson, A.R. (1984a). Student evaluation techniques and their relationship to grade and curriculum. Paper presented at the Northern Rocky Mountain Educational Research Association annual meeting, Jackson, Wyoming.

--- (1984b). Teacher perspectives of their instructional use of tests. Journal of Educational Research, 77, 244-248.

--- (1982) The practice of testing in elementary and secondary schools. (ERIC ED 229391)

McKee, B.G., & Manning-Curtis, C. (1982). Teacher-constructed classroom tests: the stepchild of measurement research. Paper presented at the National Council on Measurement in Education annual meeting, New York.

Newman, D.C., & Stallings, W.M. (1982). Teacher competency in classroom testing, measurement preparation, and classroom testing practices. Paper presented at the American Educational Research Association annual meeting, New York. (ERIC ED 220491)

Tollefson, N., Tracy, D.B., Kaiser, J., Chen, J.S., & Kleinsasser, A. (1985). Teachers' attitudes toward tests. Paper presented at the American Educational Research Association annual meeting, Chicago.

Yeh, J.P., Herman, J.L., & Rudner, L.M. (1981). Teachers and testing: a survey of test use. Center for the Study of Evaluation, California University, Los Angeles. (ERIC ED 218336)

Table 1. Use of types of items and tests by grade level taught (means and standard deviations)

Item Types	Content Area: Art/Music	English	PE	Social	Math	Science	F	p
Number of cases	29	57	26	37	42	27		
True-false	2.2 (1.1)	2.6 (.9)	3.4 (1.5)	3.0 (1.1)	2.1 (.9)	2.7 (1.0)	6.78	.01
Multiple-choice	3.2 (1.2)	3.4 (1.1)	3.3 (1.4)	4.2 (.9)	2.7 (1.1)	3.7 (.9)	7.67	.01
Completion	3.3 (1.3)	3.2 (1.2)	3.3 (1.6)	3.7 (1.1)	3.0 (1.2)	3.9 (1.2)	2.40	.04
Matching	3.3 (1.5)	3.3 (1.0)	3.0 (1.5)	3.9 (1.0)	2.6 (.7)	3.7 (1.0)	6.16	.01
Essay	2.5 (1.6)	4.0 (1.3)	2.7 (1.4)	3.8 (1.3)	1.8 (1.0)	3.6 (1.3)	18.26	.01
Short answer	3.4 (1.7)	4.0 (1.0)	3.5 (1.5)	4.4 (.9)	3.4 (1.4)	4.4 (.8)	5.46	.01
AGGREGATE	18.0 (5.3)	20.4 (3.5)	19.0 (7.0)	23.0 (3.4)	15.7 (3.5)	22.0 (3.1)	14.90	.01
Test Types								
Diagnostic	1.8 (1.0)	3.0 (1.3)	1.6 (.9)	2.8 (1.3)	3.1 (1.3)	2.3 (1.1)	8.30	.01
Norm-referenced	1.4 (.7)	1.9 (.9)	1.7 (1.1)	1.9 (1.0)	2.0 (1.0)	2.0 (1.1)	-	NS
Criterion-referenced tests--frequency of use	2.1 (1.4)	2.2 (1.3)	2.3 (1.6)	2.3 (1.4)	2.6 (1.6)	2.4 (1.4)	-	NS
-- % using	48.3%	35.2%	44.0%	48.6%	45.2%	51.9%		
Performance	3.9 (1.2)	3.1 (1.3)	4.5 (1.0)	2.4 (1.2)	3.7 (1.6)	3.0 (1.2)	9.62	.01
Competency	3.3 (1.5)	2.4 (1.2)	2.7 (1.4)	2.4 (1.3)	3.1 (1.6)	2.5 (1.3)	2.39	.04
AGGREGATE	12.1 (3.6)	12.0 (4.3)	12.4 (3.0)	11.0 (4.7)	13.8 (4.7)	12.3 (4.6)	-	NS
Sources of test items:								
Construct own	69.5% (28.3)	63.3% (24.8)	68.8% (25.8)	63.7% (26.2)	61.9% (33.1)	61.8% (28.0)	-	NS
Use manuals	22.7% (25.4)	34.7% (23.0)	29.5% (20.4)	31.4% (24.4)	39.9% (31.8)	35.4% (22.2)	-	NS
Time (hours) spent in test-related activities per week	5.2 (4.9)	8.5 (5.5)	5.5 (4.5)	6.5 (3.4)	7.1 (3.3)	7.4 (3.5)	2.59	.03
Percent of grade based on test score	27.6% (19.7)	35.3% (17.7)	30.7% (14.2)	46.2% (21.4)	49.5% (23.4)	50.1% (18.9)	7.74	.01